Adaptive Practice of Facts in Domains with Varied Prior Knowledge

Jan Papoušek
Masaryk University Brno
jan.papousek@mail.muni.cz

Radek Pelánek
Masaryk University Brno
pelanek@fi.muni.cz

Vít Stanislav
Masaryk University Brno
slaweet@mail.muni.cz

ABSTRACT

We propose a modular approach to development of a computerized adaptive practice system for learning of facts in areas with widely varying prior knowledge: decomposing the system into estimation of prior knowledge, estimation of current knowledge, and selection of questions. We describe specific realization of the system for geography learning and use data from the developed system for evaluation of different student models for knowledge estimation. We argue that variants of the Elo rating systems and Performance factor analysis are suitable for this kind of educational system, as they provide good accuracy and at the same time are easy to apply in an online system.

1. INTRODUCTION

Computerized adaptive practice [10] aims at providing students with practice in an adaptive way according to their skill, i.e. to provide the students with tasks that are most useful to them. Our aim is to make the development of such a system as automated as possible, particularly to enable the system to learn the relevant aspects of the domain from the data so that there is no need to rely on domain experts. This aspect is especially important for development of systems for small target groups of students, e.g. systems dealing with specialised topics or languages spoken by relatively small number of people (like Czech).

This work is focuses on the development of adaptive systems for learning of facts. In the terminology of the "knowledge learning instruction framework" [11] we focus on constant-constant knowledge components, i.e. knowledge components with a constant application condition and a constant response. We are particularly concerned with learning of facts in areas where students are expected to have nontrivial and highly varying prior knowledge, e.g. geography, biology (fauna, flora), human anatomy, or foreign language vocabulary. To show the usefulness of focusing on estimation of prior knowledge, Figure 1 visualizes the significant differences in prior knowledge of African countries.

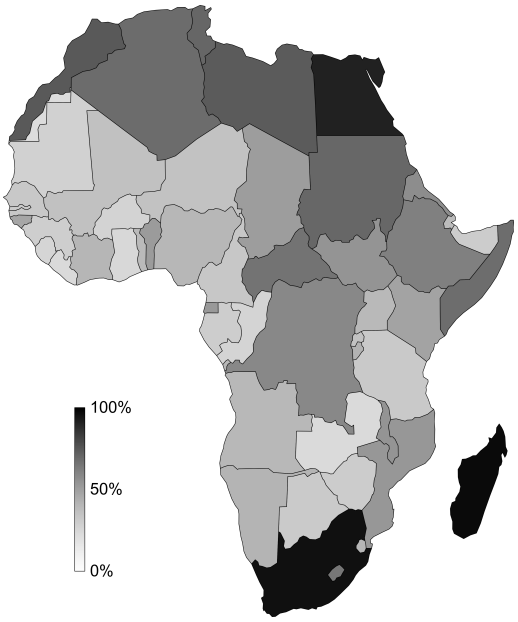

Figure 1: Map of Africa colored by prior knowledge of countries, the shade corresponds to the probability of correct answer for an average user of `slepemapy.cz`.

To achieve effective learning in domains such as geography it is necessary to address several interrelated issues, particularly the estimation of knowledge, the modeling of learning, the memory effects (spacing and forgetting), and the question selection.

The above-mentioned issues have been studied before, but separatedly in different context. Adaptation has been studied most thoroughly in the context of *computerized adaptive testing* (CAT) with the use of the item response theory [3]. In CAT the goal is the testing, i.e. to determine the skill of students. Therefore, the focus of CAT is on precision and statistical guarantees. It usually does not address learning (students' skill is not expected to change during a test) and motivation. In our setting the primary goal is to improve the skill; estimation of the skill is only a secondary goal which helps to achieve the main one. Thus the statistical accuracy of the estimation is not so fundamental as it is in CAT. On the other hand, the issues of learning, forgetting, and motivation are crucial for adaptive practice.

Another related area is the area of *intelligent tutoring systems* [23]. These systems focus mainly on learning of more complex cognitive skills than learning of facts, e.g. mathematics or physics. The modeling of learning is widely studied in this context, particularly using the popular Bayesian knowledge tracing model [2]. A lot of research focuses on the acquisition of skills, less attention is given to the prior knowledge and the forgetting (see e.g. [15, 20]).

The learning of facts is well studied in the research of *memory*, e.g. in the study of spacing and forgetting effects [16] and spaced repetition [9]. These studies are not, however, usually done in a realistic learning environment, but in a laboratory and in areas with little prior knowledge, e.g. learning of arbitrary word lists, nonsense syllables, obscure facts, or Japanese vocabulary [4, 16]. Such approach facilitates interpretation of the experimental results, but the developed models are not easily applicable in educational setting, where prior knowledge can be an important factor. There are also many implementations of the spaced repetition principle using "flashcard software" (well known example is SuperMemo), but these implementations usually use scheduling algorithms with fixed ad-hoc parameters and do not try to learn from collected data (or only in a limited way). The spaced repetition was also studied specifically for geography [26], but only in a simple setting.

In this work we propose both a general structure and a specific realization of a computerized adaptive practice system for learning of facts. We have implemented an instance of such system for learning geography, particularly names of countries (`slepemapy.cz`, the system is so far implemented only in Czech). Data from this system are used for the evaluation (over 2 500 students, 250 000 answers). To make the description more concrete and readable, we sometimes use the terminology of this system, i.e., learning of country names. Nevertheless, the approach is applicable to many similar domains (other geographical objects, anatomy, biology, foreign vocabulary).

The functionality of the system is simple: it provides series of questions about countries ("Where is country X?", "What is the name of this country?") and students answer them using an interactive map. Questions are interleaved with a feedback on the success rate and a visualization of the estimated knowledge of countries. The core of the system lies in estimating students' knowledge and selecting suitable questions.

We decompose the design of such system into three steps and treat each of these steps independently:

1. *Estimation of prior knowledge.* Estimating the probability that a student s knows a country c before the first question about this country. The estimate is based on previous answers of the student s and on answers of other students about the country c.

2. *Estimation of current knowledge.* Estimating the probability that the student s knows a country c based on the estimation of prior knowledge and a sequence of previous answers of student s on question about country c.

3. *Selection of question.* Selection of a suitable question for a student based on the estimation of knowledge and the recent history of answers.

Each of these issues is described and evaluated in a single section. The independent treatment of these steps is a useful simplifications, since it makes the development of the system and student models more tractable. Nevertheless, it is clearly a simplification and we discuss limitations of this approach in the final section.

2. BACKGROUND

In this section we briefly describe some of the relevant models that are used in the realization and evaluation of our approach.

2.1 Bayesian Knowledge Tracing

Bayesian knowledge tracing (BKT) [2, 21] is a well-known model for modeling of learning (changing skill). It is a hidden Markov model where skill is the binary latent variable (either learned or unlearned). The model has 4 parameters[1]: probability that the skill is initially learned, probability of learning a skill in one step, probability of incorrect answer when the skill is learned (slip), and probability of correct answer when the skill is unlearned (guess). The skill estimated is updated using a Bayes rule based on the observed answers. Parameter estimation can be done using the Expectation Maximization algorithm or using the exhaustive search.

2.2 Rasch Model

Basic model in the item response theory is the Rasch model (one parameter logistic model). This model assumes the student's knowledge is constant and expressed by a skill parameter θ, the item's difficulty is expressed by a parameter b, and the probability of a correct answer is given by the logistic function:

$$P(correct|b,\theta) = \frac{1}{1+e^{-(\theta-b)}}$$

The standard way to estimate the parameters from data is to use the joint maximum likelihood estimation [3], which is an iterative procedure. In the case of multiple choice question with n options, the model is modified to use a shifted logistic function:

$$P(correct|b,\theta) = \frac{1}{n} + (1-\frac{1}{n})\frac{1}{1+e^{-(\theta-b)}}$$

2.3 Performance Factor Analysis

Performance factor analysis (PFA) [17] can be seen as an extension of Rasch model with changing skill. The skill, which is a logit of probability of a correct answer, is given by a linear combination of the item's difficulty and the past successes and failures of a student:

$$P(correct) = \frac{1}{1+e^{-m}}$$

$$m = \beta + \gamma s + \delta f$$

[1]BKT can also include forgetting. The described version corresponds to the variant of BKT that is most often used in research papers.

where β is the item difficulty, s and f are counts of previous successes and failures of the student, γ and δ are parameters that determine the change of the skill associated with correct and incorrect answer. Note that originally PFA [17] is formulated in terms of vectors, as it uses multiple knowledge components; for our analysis the one-dimensional version is sufficient.

2.4 Elo System

The Elo rating system [5] was originally devised for chess rating, i.e. estimating players skills based on results of matches. For each player i we have an estimate θ_i of his skill, based on the result R (0 = loss, 1 = win) of a match with another player j; the skill estimate is updated as follows:

$$\theta_i := \theta_i + K(R - P(R = 1))$$

where $P(R = 1)$ is the expected probability of winning given by the logistic function with respect to the difference in estimated skills, i.e. $P(R = 1) = 1/(1 + e^{-(\theta_i - \theta_j)})$, and K is a constant specifying sensitivity of the estimate to the last attempt. An intuitive improvement, which is used in most Elo extensions, is to use an "uncertainty function" instead of a constant K. There are several extension to the Elo system in this direction, the most well-known is Glicko [6].

We can use the Elo system in student modeling, if we interpret a student's answer on an item as a "match" between the student and the item. Recently, several researchers have studied this kind of application of the Elo system in the educational data mining [10, 24, 25].

The basic Elo system (reinterpreted in the context of educational problems) also uses the logistic function and one parameter for each student and problem. Thus the Rasch model and the Elo system are in fact very similar models, the main principal difference is that the Rasch model assumes the constancy of parameters, the Elo system assumes a changing skill.

3. ESTIMATION OF PRIOR KNOWLEDGE

At first, we treat the estimation of prior knowledge. Our aim is to estimate the probability that a student s knows a country c based on previous answers of students s to questions about different countries and previous answers of other students to questions about country c – as a simplification (for an easier interpretation of data) we use only the first answer about each country for each student in this step.

3.1 Model

In the following text we use a key assumption that both students and studied facts are homogenous; we assume that we can model students' overall prior knowledge in the domain by a one-dimensional parameter. This assumption is reasonable for geography and students from Czech Republic (which is the case of our application), but would not hold for geography and mixed population or for a mix of facts from geography and chemistry. If the homogenity is not satisfied, we can group the students and facts into homogenous groups (e.g. students by their IP address, facts by an expert or by an automatic technique [1]) and then make predictions for each subgroup independently.

More specifically, we model the prior knowledge by the Rasch model, i.e. we have student parameter θ_s corresponding to the global knowledge of a student s of geography, the item parameter b_c corresponding to the difficulty of a country c, and the probability of a correct first answer is given by the logistic function $P(\mathit{correct}|s, c) = \frac{1}{1+e^{-(\theta_s - b_c)}}$.

As we mentioned above, the standard approach to the parameter estimation for the Rasch model is joint maximum likelihood estimation (JMLE). This is an iterative approach that is slow for large data, particularly it is not suitable for an online application, where we need to adjust estimates of parameters continuously.

Therefore, we also consider the application of the Elo rating system in this setting. Although the assumptions in this context are closer to the assumptions of the Rasch model (the global skill and the difficulty of items are rather constant), the Elo system is much more suitable for an online application and results with simulated data suggest that it leads to similar estimates [19].

3.2 Evaluation

The basic version of the Elo system with the constant update parameter K does not provide a good estimation – if the parameter K is small, the system takes long to learn skills and difficulties, if the parameter K is large, the behavior of the system is unstable (estimates are too dependent on a last few answers). Therefore, instead of the constant K we use an uncertainty function $\frac{a}{1+bn}$, where n is the order of the answer and a, b are parameters. Using a grid search we have determined optimal values $a = 1, b = 0.05$. This exact choice of parameter values is not important, many different choices of a, b provide very similar results.

This variant of the Elo system provides both fast coarse estimates after a few answers and stability in the long run (see Figure 2 A). It also provides nearly identical estimates as the joint maximum likelihood estimation (Figure 2 B, correlation 0.97). JMLE is computationally demanding iterative procedure, the Elo system requires a single pass of the data and can be easily used online. Since the estimates of the two methods are nearly identical, we conclude that the Elo system is preferable in our context.

Distribution of the difficulty parameters (Figure 2 C) reflects the target domain and student population. In our case the difficulty of countries for Czech students is skewed towards very easy items, which are mostly European countries. Difficult countries are mostly African. Skill parameters are distributed approximately normally.

We have tested the assumption of a single global skill by computing the skill for independent subsets of items (countries from different continents) and then checking the correlation between the obtained skill. Figure 2 D shows the results for two such particular "subskills", the correlation coefficient for this case and other similar pairs of subskills is around 0.6. Given that there is some intrinsic noise in the data and that the skills are estimated from limited amount of questions, this is quite high correlation. This suggests that the assumption of a global skill is reasonable.

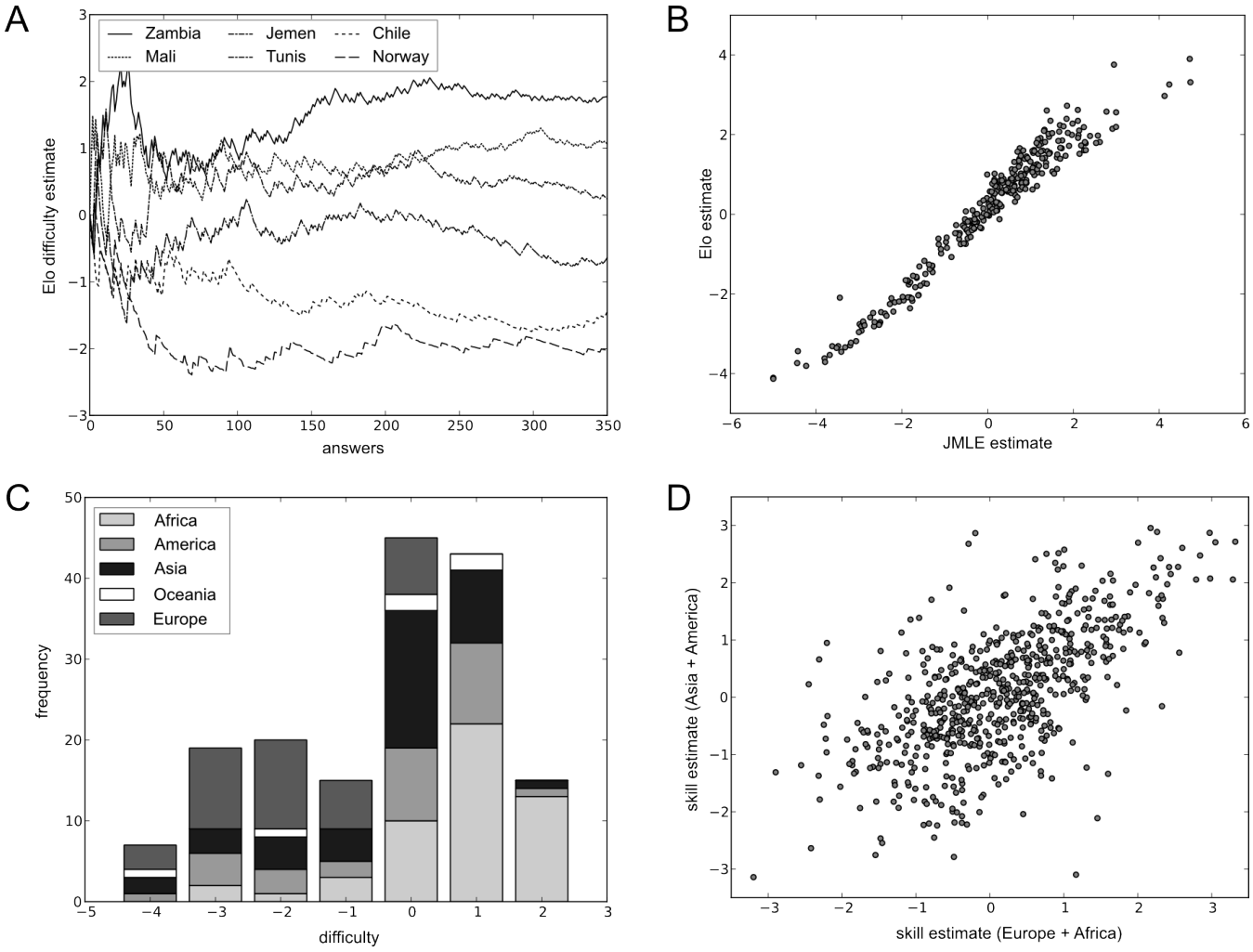

Figure 2: Estimation of prior knowledge: A) Development of estimates of difficulty of selected countries under Elo system, B) Comparison of Elo and JMLE difficulty estimates, C) Histogram of difficulty of countries, D) Correlation of "subskills" computed for different sets of countries.

4. ESTIMATION OF CURRENT KNOWLEDGE

We now turn to the estimation of a student's current knowledge, i.e. knowledge influenced by the repeatedly answering of questions about a country. The input data for this estimation are an estimate of prior knowledge (provided by the above described model) and the history of previous attempts, i.e. the sequence of previous answers (correctness of answers, question types, timing information).

4.1 Models

Several different models can be considered for the estimation of current knowledge. Bayesian knowledge tracing can be used in a straightforward way. In this context the probability of initial knowledge is given by the previous step. The probability of learning, guess, and slip are either given by a context (guess in the case of multiple choice question) or can be easily estimated using an exhaustive search. However, in this context the assumptions of BKT are not very plausible. BKT assumes a discrete transition from the unknown to the known state, which may be reasonable a simplification for procedural skills, but for declarative facts the development of the memory is gradual.

Assumptions of the Performance factor analysis are more relevant for the learning of facts. Instead of the item difficulty parameter β_i, used in the original version of PFA, we can use the estimate of the initial knowledge for a student s and a country c in our setting. This is given by the difference $\theta_s - b_c$.

A disadvantage of PFA is that it does not consider the order of answers (it uses only the summary number of correct and incorrect answers) and it also does not take into account the probability of guessing. Guessing can be important particularly in our setting, where the system uses multiple choice questions with variable number of options. To address these issues we propose to combine PFA with some aspects of the Elo system (in the following text we denote this version as PFAE – PFA Elo/Extended):

- K_{sc} is the estimated knowledge of a student s of a country c.
- The initial value of K_{sc} is provided by the estimation of prior knowledge: $K_{sc} = \theta_s - b_c$.
- The probability of correct answer to a question with n options is given by the shifted logistic function:

$$P(correct|K_{sc}, n) = \frac{1}{n} + (1 - \frac{1}{n})\frac{1}{1 + e^{-K_{sc}}}$$

- After a question with n options was answered, the estimated knowledge is updated as follows:
  - $K_{sc} := K_{sc} + \gamma \cdot (1 - P(correct|K_{sc}, n))$, if the answer was correct,
  - $K_{sc} := K_{sc} + \delta \cdot P(correct|K_{sc}, n)$, if the answer was incorrect.

The estimation can be further improved by taking into account the timing information. If two questions about the same item are asked closely one after another, then it can be expected that the student will answer the second one correctly, because the answer is still in his short term memory. In models based on a logistic function (PFA, PFAE) we can model this effect in the following way: the skill is "locally" increased by $\frac{w}{t}$, where t is the time (in seconds) between attempts and k is a suitable constant (optimal $w = 80$ for our data). It should be possible to further improve the model by a more thorough treatment of forgetting and spacing effects, e.g., by incorporating some aspects of the ACT-R model [16].

Another useful timing information is the response time. As the response time tends to be log-normally distributed [8, 22], we work with the logarithm of time. Intuitively, the higher knowledge of a country leads not only to higher probability of a correct answer, but also to a faster response. Figure 3 shows results of an experiment supporting this intuition – distribution of times of correct answers is shifted to lower values if the next answer on the same country is correct. This suggests that response time could be used to improve the estimation of knowledge. Indeed, even simple modification of the γ parameter in the PFA model (by comparison of the response time to mean response time) leads to a slight improvement in predictions. A more involved application of the response time requires a suitable normalization due to different speeds of students and different sizes of countries – it is much easier to click on China than on Vietnam.

4.2 Evaluation

The described models provide predictions of probability of a correct answer. To evaluate these models we need to choose a metric by which we measure performance of models. In educational data mining researchers often provide evaluation with respect to a chosen metric without providing any rationale for the particular choice. In some cases the choice of metric is not fundamental and different metrics lead to similar results (that is the case for above described experiments with estimating prior knowledge). However, the evaluation of models of the current knowledge is sensitive to the choice of a metric, and thus it is necessary to pay attention to this issue.

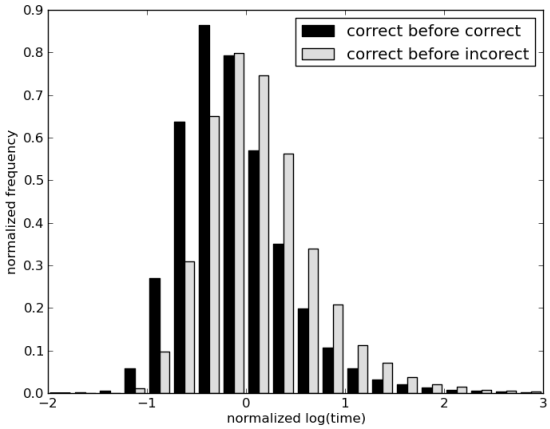

Figure 3: Normalized logarithm of time of correct answers, depending on whether the next answer about the country is answered correctly or incorrectly.

Let us review the most commonly used metrics in educational data mining and their suitability in our context. The mean absolute error (MAE) is not a good metric, since for unbalanced data it prefers models skewed towards the larger class. Consider a simulated student that answers correctly with constant probability 0.7. If we optimize a constant predictor with respect to the mean absolute error, the predicted probability is 1. The root mean square error (RMSE) is a similar measure that does not share this disadvantage and is thus preferable. The log-likelihood (LL) metric behaves similarly to RMSE except for predictions very close to 0 or 1. Since LL is unbounded, a single wrong prediction can degrade the performance of a model. To prevent this behaviour, an ad-hoc bound can be introduced in the computation of LL. Metrics like AIC and BIC are extensions of the log-likelihood penalizing large number of model parameters. All models described above have only very small number of parameters, and thus these metrics are not relevant for the current discussion. Another popular metric is the area under the receiver operating characteristic curve (AUC). This metric considers the prediction only in relative way – note that if all predictions are divided by 2, the AUC metric stays the same. In our application, however, the precision of the absolute prediction is important, since the value is used in computations that determine the choice of questions and number of options in multiple choice questions.

Thus it seems that the most suitable metrics from the commonly used ones is RMSE. Thus we use RMSE as our primary metric, i.e. to optimize values of model parameters. Table 1 provides a comparison of different models also for other metrics. We can see that the results are inconclusive regarding the comparison of BKT and PFA, but the newly proposed extension PFAE beats both the standard PFA and BKT models with respect to all three reported metrics. The results also show that the consideration of timing information further improves the performance of models.

Table 1: Model comparison.

model	RMSE	LL	AUC
BKT	0.262	-42048	0.668
PFA	0.265	-44740	0.669
PFA + time	0.262	-43088	0.695
PFAE	0.262	-41947	0.682
PFAE + time	0.259	-40623	0.714

For the reported evaluation we use models with "global" parameters, i.e., for example in the PFA and its extension we use the same parameters γ, δ for all countries and students. Thus the models have very small number of parameters (at most 4 for the extension with timing information) and can be easily fit by an exhaustive search. Since the number of data points is many orders larger (tens of thousands), overfitting is not an issue. It would be possible to use the "local" parameter values for individual countries and students, such variant would require an improved parameter estimation and a mechanism for dealing with uneven distribution of data among countries and students.

5. QUESTION SELECTION

We will now focus on the issue of the question selection. Based on the past performance of the student we want to select a suitable next question. In the context of our geography application the selection of a question consists of several partial decisions: which country to target, which type of the question to ask ("Where is X?" versus "What is the name of this country?"), and how many options to give a student to choose from.

Compared to the knowledge estimation, the question selection is much harder to evaluate, since we do not have a single, clear, easily measurable goal. The overall goal of the question selection is quite clear – it is the maximization of student learning. But it is not easy to measure the fulfilment of this general goal, since it depends also on the context of the learning. An experiment with pre-test, post-test, and fixed time in the system may provide a setting for an accurate evaluation of the different question selection strategies. Results of such experiment would, however, lack ecological validity, as many of the users of the system use the system on their own and with variable time in the system, so for example the issue of motivation is much more important than in a controlled experiment. A related work [18] presents this kind of controlled experiment for card selection in drill practice, the authors however provide comparison only with respect to a very simple cyclic selection technique and not to an evaluation of different alternatives of the selection algorithm. Another possibility is to use the time spent in educational system as a measure of quality of question selection. Here, however, the optimal choice with respect to this measure may not be optimal for learning, see [12] for a specific instance of an educational online game with this dynamics.

Thus at the moment we do not provide the evaluation of the question selection. We formulate general criteria that the question selection should satisfy and propose a specific approach to achieve these criteria.

5.1 Criteria

The question selection process should satisfy several criteria, which are partly conflicting. The criteria and their weight may depend on the particular application, the target student population, and student goals. We propose the following main criteria.

The selection of question should depend on an estimated *difficulty* of question. From the testing perspective, it is optimal to use questions with expected probability of a correct answer reaching 50%, because such question provide most information about students' knowledge. However, 50% success rate is rather low and for most students it would decrease motivation. Thus in our setting (adaptive practice) it is better to aim for a higher success rate. At the moment we aim at 75%, similarly to previous work [7].

Another important issue is the *repetition* of questions. This aspect should be governed by the research about spacing effects [4, 16], particularly it is not sensible to repeat the same question too early.

It may be also welcome to have *variability* of question types. Different question types are useful mainly as a tool for tuning the difficulty of questions, but even if this is not necessary, the variability of question types may be meaningful criteria in itself, since it improves user experience, if used correctly.

5.2 Selecting Target Country

We propose to use the linear scoring approach to select a target country (the correct answer of the question). For each relevant attribute, we consider a scoring function that expresses the desirability of a given country with respect to this attribute. These scoring functions are combined using weighted sum, the country with highest total score is selected as a target. We consider the following attributes:

1. the probability the student knows the country,
2. time since the last question about the country,
3. the number of questions already answered by the student about the country.

Figure 4 shows the general shape of scoring functions for these attributes. Further we specify concrete formulas that approximate these shapes using simple mathematical functions.

The first case takes into account the relation between the estimated probability of a correct answer (P_{est}) and the target success rate (P_{target}). Assume that our goal is to ask a question where the student has 75% chance of a correct answer. The distance from the probability for the difficult countries (nearly 0% chance of the correct answer) is higher than for easy ones (almost 100%), so it is necessary to normalize it.

$$S_{prob}(P_{est}, P_{target}) = \begin{cases} \frac{P_{est}}{P_{target}} & \text{if } P_{target} \geq P_{est} \\ \frac{1-P_{est}}{1-P_{target}} & \text{if } P_{target} < P_{est} \end{cases}$$

The second scoring function penalizes countries according to the time elapsed since the last question, because we do not want to repeat countries in a short time interval when

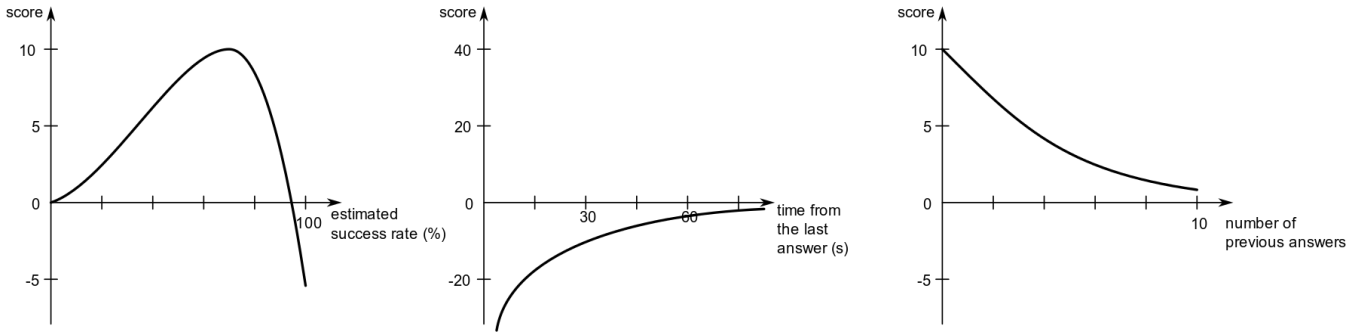

Figure 4: Desired contribution of different criteria to selection of target country.

they are still in short term memory. We use the function $S_{time}(t) = -1/t$, where t is time in seconds. Using just the above mentioned attributes the system would ask questions for only a limited pool of countries. To induce the system to ask questions about new countries we introduce the third scoring function that uses the total number n of questions for the given country answered by the student: $S_{count}(n) = 1/\sqrt{1+n}$. The total score is given as a weighted sum of individual scores, the weights are currently set manually, reflecting experiences with the system: $W_{prob} = 10$, $W_{count} = 10$, $W_{time} = 120$.

5.3 Choosing Options

Once the question's target is selected, the question can be adjusted according to the student's needs by using a multiple choice question with suitable number of options. For a multiple choice question the probability of a correct answer is the combination of the probability of guessing the answer (P_{guess}) and knowing the target country (P_{est})[2]:

$$P_{success} = P_{guess} + (1 - P_{guess}) \cdot P_{est}$$

As our goal is to get $P_{success}$ close to P_{target}, we would like to make P_{guess} close to

$$G = \frac{P_{target} - P_{est}}{1 - P_{est}}$$

For $G \leq 0$, we use open question (no options), otherwise we use n closest to $\frac{1}{G}$ as a number of options. For principal reasons the minimal possible value of n is 2, for practical reasons there is also an upper bound for n (more than 6 options would be confusing). The type of the question – "Where is country X?" or "What is the name of this country?" is currently selected randomly. In case of an open question the first type is always used.

When using multiple choice questions, we also need to choose the distractor options. Unlike other systems for practice dealing with text [13, 14], we work with well structured data, so the problem of option selection is easier. The choice of options can be based on domain information, e.g. geographically close countries or countries with similar names. The easiest way to choose good distractors is, however, to simply base the choice on past answers. We can take countries most commonly mistaken with the target country (in open questions) and select from them randomly. The random choice is weighted by the frequency of mistakes with the given country, for example Kamerun is most often confused with Niger (38%), Nigeria (27%), Central African Republic (10%), Republic of the Congo (9%), Gabon (6%), Ivory Coast (5%), Uganda (3%), and Guinea (2%).

[2]This is, of course, simplification since a multiple choice question can also be answered by ruling out distractor options. But if the distractors are well chosen; this simplification is reasonable.

6. DISCUSSION

We described the functionality of the system in three independent parts: the estimation of prior knowledge, the estimation of current knowledge, and the selection of a question. The independent treatment of these steps is, however, a simplification, as there is an interaction between these steps.

In our treatment, only the first answer about a given item is taken as an indication of a prior knowledge, other answers are considered as an indication of changes in knowledge. But for example the second answer, clearly, also contains some information about prior knowledge. A more precise models should be possible by incorporating more integrated approach to the estimation of prior and current knowledge.

The selection of a question was treated as a subsequent step after the estimation of knowledge, but in reality there is a feedback loop: the estimation of knowledge influences the selection of a question and the selection of a question determines the data that are collected and used for the estimation of knowledge. Since the collected data are partially determined by the model used, there may be a bias in the data towards certain questions, and this bias may, in a subtle way, influence the evaluation. For example, if the model overestimates the knowledge of students, the question selection stops asking questions about items too early, which means that the system does not collect data that would contradict the overestimated knowledge. The question selection procedure may be also modified in such a way to collect data most useful for improving the precision of the estimation. The study of these interactions may be more important than differences between different models or estimation procedures, which typically get most attention in current research in student modeling.

Acknowledgements

Authors thank anonymous reviewers for valuable comments, Jiří Říhák and Juraj Nižnan for fruitful discussions, and Tereza Doležalová for language assistance.

7. REFERENCES

[1] P. Boroš, J. Nižnan, R. Pelánek, and J. Řihák. Automatic detection of concepts from problem solving times. In *Proc. of International Conference on Artificial Intelligence in Education (AIED 2013)*, volume 7926 of *LNCS*, pages 595–598. Springer, 2013.

[2] A. Corbett and J. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[3] R. De Ayala. *The theory and practice of item response theory*. The Guilford Press, 2008.

[4] P. F. Delaney, P. P. Verkoeijen, and A. Spirgel. Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of learning and motivation*, 53:63–147, 2010.

[5] A. E. Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[6] M. E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999.

[7] B. R. Jansen, J. Louwerse, M. Straatemeier, S. H. Van der Ven, S. Klinkenberg, and H. L. Van der Maas. The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, 24:190–197, 2013.

[8] P. Jarušek and R. Pelánek. Analysis of a simple model of problem solving times. In *Proc. of Intelligent Tutoring Systems (ITS)*, volume 7315 of *LNCS*, pages 379–388. Springer, 2012.

[9] J. D. Karpicke and H. L. Roediger. Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57(2):151–162, 2007.

[10] S. Klinkenberg, M. Straatemeier, and H. Van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011.

[11] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012.

[12] D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger. Optimizing challenge in an educational game using large-scale design experiments. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 89–98. ACM, 2013.

[13] R. Mitkov, L. A. Ha, and N. Karamanis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194, 2006.

[14] J. Mostow, B. Tobin, and A. Cuneo. Automated comprehension assessment in a reading tutor. In *ITS 2002 Workshop on Creating Valid Diagnostic Assessments*, pages 52–63, 2002.

[15] Z. A. Pardos and N. T. Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*, pages 255–266. Springer, 2010.

[16] P. I. Pavlik and J. R. Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29(4):559–586, 2005.

[17] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press, 2009.

[18] P. Pavlik Jr, T. Bolster, S.-M. Wu, K. Koedinger, and B. Macwhinney. Using optimally selected drill practice to train basic facts. In *Intelligent Tutoring Systems*, pages 593–602. Springer, 2008.

[19] R. Pelánek. Time decay functions and elo system in student modeling. In *Proc. of Educational Data Mining (EDM)*, 2014.

[20] Y. Qiu, Y. Qi, H. Lu, Z. A. Pardos, and N. T. Heffernan. Does time matter? modeling the effect of time with bayesian knowledge tracing. In *Proc. of Educational Data Mining (EDM)*, pages 139–148, 2011.

[21] B. van de Sande. Properties of the bayesian knowledge tracing model. *Journal of Educational Data Mining*, 5(2):1, 2013.

[22] W. Van der Linden. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2):181, 2006.

[23] K. Vanlehn. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265, 2006.

[24] K. Wauters, P. Desmet, and W. Van Den Noortgate. Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6):549–562, 2010.

[25] K. Wauters, P. Desmet, and W. Van Den Noortgate. Monitoring learners' proficiency: Weight adaptation in the elo rating system. In *EDM*, pages 247–252, 2011.

[26] D. M. Zirkle and A. K. Ellis. Effects of spaced repetition on long-term map knowledge recall. *Journal of Geography*, 109(5):201–206, 2010.